Appendix

Table of Contents

A Broader impacts

The authors discussed potential negative impacts among ourselves. The primary potential negative impact we identified is that we raise awareness of datasets and pre-trained models derived from retracted datasets or released under licenses not adherent to the intent of the license of the parent. However, many of these datasets and pre-trained models are already widely used and accessible, so we do not believe our documentation will cause much additional harm. Instead, we hope that our work makes ethical considerations more clear to users of these assets.

B Supplemental data

Supplemental data is available at `https://github.com/citp/mitigating-dataset-harms`.

Maintenance. This data will remain available indefinitely as long as the Princeton CITP GitHub is operational. Links used to access publicly-available PDFs may eventually deprecate, but the DOIs we give ensure that all papers included in our analysis remain identifiable.

License. License details for our data can be found at the above link: `https://github.com/citp/mitigating-dataset-harms`.

Available files. We make the following `.csv` files available:

`msceleb1m_all.csv, dukemtmc_all.csv, lfw_all.csv` These are the full corpora we collected, containing 1,404, 1,393, and 7,732 papers respectively. The following columns are given, and reflect information given by Semantic Scholar:

- **paperId:** the Semantic Scholar id of the paper
- **cites {dataset id}:** for each dataset used to build the corpus, 1 if the paper cites {dataset id} and 0 otherwise—see Table 4 for dataset ids.

- **title, abstract, year, venue, arxivId, doi**
- **pdfUrl:** a URL where the paper may be publicly available, found via Semantic Scholar or arXiv

`msceleb1m_labeled.csv, dukemtmc_labeled.csv, lfw_labeled.csv` These are the samples of papers that we analyzed, containing 276, 275, and 400 papers respectively. In addition to all the columns above, the following additional columns are given:

- **uses dataset or derivative:** 1 if we determined that the paper uses a dataset or derivative and 0 otherwise
- **dataset(s) / model(s) used:** a comma separated list of datasets or models used, denoted by the id provided in Table 4 in brackets (e.g., [D8], [M5])
- **unable to disambiguate:** 1 if we were unable to determine the specific dataset(s) used or whether a dataset was used, and 0 otherwise

`summary.csv` An extended version of Table 4.

`dataset_list.csv` This file contains the names of the 54 face and person recognition datasets we compiled to select our three datasets of interest, the number of total citations on Semantic Scholar (at the time of collection in August 2020), and their Semantic Scholar Corpus ID which can be used to access metadata from the Semantic Scholar API [34].

This list was compiled from datasets mentioned in the following six sources [23, 94, 55, 84, 45, 38] and cited at least 100 times. Sorted from most total citations to least, the datasets are: Yale Face Database B, LFW, FERRET, AR, VGGFace, LFWA, CelebA, AU-Coded Facial Expression Database (Cohn-Kanade), FRGCv2, Extended Yale Face Database B, CK+, JAFFE, PIE, Multi-PIE, RaFD, PubFig, CelebFaces+, XM2VTS, BU-3DFE, CASIA-WebFace, YTF, MMI, MORPH, DukeMTMC, MS-Celeb-1M, Bosphorus, VGGFace2, CUFS, Replay-Attack, IJB-A, YTC, Attributes 25K, MegaFace, FER-2013, FaceTracer, CASIA-FASD, Berkeley Human Attributes, Oulu-CASIA, AffectNet, Brainwash, CACD, EmotioNet, SPEW 2.0, PaSC, CUFSF, CFP, CASIA NIR-VIS v2.0, RAF-DB, IJB-C, IJB-B, Oxford TownCentre, CASIA-HFB, UMDFaces, and Ego-Humans.

C Description of datasets analyzed

In this section, we provide details about the three datasets our analysis focused on: MS-Celeb-1M, DukeMTMC, and Labeled Faces in the Wild.

MS-Celeb-1M was introduced by Microsoft researchers in 2016 as a face recognition dataset [41]. It includes about 10 million images of about 10,000 "celebrities." The original paper gave no specific motivating applications, but did note that "Recognizing celebrities, rather than a pre-selected private group of people, represents public interest and could be directly applied to a wide range of real scenarios." Researchers and journalists noted in 2019 that many of the "celebrities" were in fact fairly ordinary citizens, and that the images were aggregated without consent [45, 66]. Several corporations tied to mass surveillance operations were also found to use the dataset in research papers [45, 66]. The dataset was taken down in April 2019. Microsoft, in a statement to the Financial Times, said that the reason was "because the research challenge is over." [66]

DukeMTMC was introduced in 2016 as a benchmark for evaluating multi-target multi-camera tracking systems, which "automatically track multiple people through a network of cameras." [72] The dataset's creators defined performance measures aimed at applications where preserving identity is important, such as "sports, security, or surveillance." The images were collected from video footage taken on Duke University's campus. The same reports on MS-Celeb-1M listed above [45, 66] noted that the DukeMTMC was also being used by corporations tied to mass surveillance operations, and also noted the lack of consent given by people included in the dataset. The creators removed the dataset in April 2019, and subsequently apologized, noting that they had inadvertently broken guidelines provided by the Duke University IRB.

LFW was introduced in 2007 as a benchmark dataset for face verification [49]. It was one of the first face recognition datasets that included faces from an unconstrained "in-the-wild" setting, using faces scraped from Yahoo News articles (via the Faces in the Wild dataset [15]). In the originally-released

paper, the dataset's creators gave no motivating applications or intended uses beyond studying face recognition. In fall 2019, a disclaimer was added to the dataset's associated website, noting that the dataset should not be used to "conclude that an algorithm is suitable for any commercial purpose." [2]

D Methodology for overarching analysis

We started our analysis of DukeMTMC, MS-Celeb-1M, and LFW by using the Semantic Scholar API [34] to record all papers citing their associated papers. Because papers that used derivatives may not always cite the original dataset, we also aimed to pull papers citing associated papers of derivatives. We identified these in a semi-automated fashion: From the list of papers above, we first downloaded PDF versions when they were publicly available either through arXiv or via links provided by Semantic Scholar. We used GROBID [5] to parse these PDFs into plaintext. We then pulled short excerpts containing keywords related to the parent dataset, which we identified through a preliminary review of papers using the dataset. By manually analyzing these excerpts, we identified derivatives that contained these keywords. We further analyzed a sample of papers to identify additional derivatives that did not contain these keywords. We retained derivatives that were cited at least 100 times to build a corpus of papers.

We combined the three parents datasets with these compiled derivatives, and recorded all the papers that cited these datasets, again using the Semantic Scholar API. The resulting corpora for DukeMTMC, MS-Celeb-1M, and LFW contained 1,393, 1,404, and 7,732 papers respectively. These corpora—assembled in December 2020—contain a large subset of papers using the three parent datasets and their derivatives. However, the corpora do not include all papers using the parent datasets and their derivatives. There are a few reasons for this. Our corpora only includes papers added to Semantic Scholar by December 2020, and Semantic Scholar itself does not index all papers.[5] Some papers are also missing because the list of derivatives we used to build each corpus is not complete. This means that the results presented throughout our paper are underestimates.

Since these were a large number of papers to examine manually, we sampled 20% or 400 papers (whichever was fewer) stratified over the year of publication.[6] In total, our analysis included 946 unique papers: 275 citing DukeMTMC or its derivatives, 276 citing MS-Celeb-1M or its derivatives, and 400 citing LFW or its derivatives. The first author coded these papers, recording whether a paper used the parent dataset or a derivative as well as the name of the parent dataset or derivative. If the first author was unable to determine the specific dataset used or whether a dataset was used, he recorded this information. A few example cases that were difficult to disambiguate are shown in Table 7.

A summary of our overarching analysis is given in Table 4.

E Supplement: Retractions of DukeMTMC and MS-Celeb-1M

We describe in detail our findings summarized in Table 2 about the retractions of MS-Celeb-1M and DukeMTMC.

Continued availability. Despite their retractions in April 2019, data from MS-Celeb-1M and DukeMTMC remain publicly available. Five of the seven derivatives of DukeMTMC either contained subsets of or the entire original dataset. The two most popular derivatives—DukeMTMC-ReID [95] and DukeMTMC-VideoReID [88]—are still available for download. Both DukeMTMC-ReID and DukeMTMC-VideoReID contain a cropped and edited subset of the videos from DukeMTMC.

Similarly, six derivatives of MS-Celeb-1M contained subsets of or the entire original dataset. Four of these—MS1MV2 [30], MS1M-RetinaFace [31], MS1M-IBUG [29], and MS-Celeb-1M-v1c [4]—are

[4]The dataset itself is no longer available. However, a script to convert DukeMTMC-ReID (which is still available) to Occluded-DukeMTMC remains available.

[5]We reproduced our corpora in August 2021, and found small discrepancies compared to our corpora collected in December 2020. The number of 2020 papers in our initial corpora for MS-Celeb-1M, DukeMTMC, and LFW are 8% less, 3% less, and 1% more, respectively, compared to the August 2021 version.

[6]We did not consider years with fewer than 10 papers. We only considered academic papers (both preprint and publications), ignoring other articles like dissertations and textbooks.

Table 4: Summary of our overarching analysis.

Dataset id	Dataset name	Associated paper	Dataset or model	Assoc. paper sampled	Num. sampled	Doc. uses	2020 doc. uses	New application	Attribute annotations	Post-processing	Still available	Includes orig. imgs.	Prohibits comm. use
D1	DukeMTMC	[72]	dataset	✓	164	14	1					✓	✓
D2	DukeMTMC-ReID	[95]	dataset	✓	172	142	63	✓			✓	✓	✓
D3	DukeMTMC-VideoReID	[88]	dataset	✓	24	11	5	✓			✓	✓	✓
D4	DukeMTMC-Attribute	[58]	dataset			10	1	✓	✓		✓		✓
D5	DukeMTMC4ReID	[37]	dataset			3	0	✓				✓	✓
D6*	DukeMTMC Group	[89]	dataset			3	1	✓					
D7	DukeMTMC-SI-Tracklet	[54]	dataset			1	1	✓				✓	✓
D8	Occluded-DukeMTMC	[64]	dataset			1	1	✓			✓[4]	✓	✓
D9	MS-Celeb-1M	[41]	dataset	✓	153	41	11				✓	✓	✓
D10	MS1MV2	[30]	dataset	✓	183	13	8			✓	✓	✓	
D11	MS1M-RetinaFace	[31]	dataset			2	2			✓	✓	✓	
D12	MS1M-LightCNN	[87]	dataset			3	0			✓	✓		
D13	MS1M-IBUG	[29]	dataset			3	1			✓	✓	✓	
D14	MS-Celeb-1M-v1c	[4]	dataset			6	4			✓	✓	✓	✓
D15	RFW	[83]	dataset			1	1	✓	✓		✓	✓	✓
D16	MS-Celeb-1M lowshot	[41]	dataset			4	0					✓	✓
D17*	Universe	[8]	dataset			2	1						
M1	VGGFace		model			6	3	✓			✓		
M2	Prob. Face Embeddings		model			1	1	✓			✓		
M3	ArcFace / InsightFace		model			14	13	✓			✓		some
M4	LightCNN		model			4	3	✓			✓		some
M5	FaceNet		model			12	5	✓			✓		
M6	DREAM		model			1	1	✓			✓		
D18	LFW	[49]	dataset	✓	220	105					✓	✓	
D19	LFWA	[59]	dataset	✓	158	2		✓	✓		✓	✓	
D20	LFW-a	[86]	dataset	✓	31	14				✓	✓	✓	
D21	LFW3D	[46]	dataset	✓	24	3				✓	✓	✓	
D22	LFW deep funneled	[50]	dataset	✓	18	4				✓	✓	✓	
D23	LFW crop	[73]	dataset	✓	8	2				✓	✓	✓	
D24	BLUFR protocol	[57]	dataset	✓	2	1					✓		
D25*	LFW87	[56]	dataset	✓	7	1							
D26	LFW+	[42]	dataset	✓	12	0		✓	✓		✓	✓	✓
D27	<no name given>	[53]	dataset			4		✓	✓		✓		
D28	<no name given>	[40]	dataset			4					✓		
D29	SMFRD	[92]	dataset			1		✓			✓	✓	
D30	LFW funneled	[48]	dataset			2				✓	✓	✓	
D31	<no name given>	[1]	dataset			2		✓	✓				
D32	<no name given>	[16]	dataset			1					✓		
D33	MTFL	[93]	dataset			1		✓	✓		✓	✓	
D34	PubFig83 + LFW	[10]	dataset			2					✓	✓	
D35	Front. Faces in the Wild	[33]	dataset			1					✓	✓	
D36	ITWE	[96]	dataset			1		✓			✓	✓	✓
D37	Extended LFW	[79]	dataset			2					✓	✓	
D38	<no name given>	[27]	dataset			1							✓

Condensed key for Table 4. *assoc. paper sampled* — yes if our corpus included a sample of papers citing the dataset's associated paper(s); *doc. uses* — the number of uses of the dataset that we were able to document; *new application* — if the derivative explicitly or implicitly enables a new application that can raise ethical questions; *attribute annotation* — if the derivative includes labels for sensitive attributes such as race or gender; *post-processing* — if the derivative manipulates the original images (for example, by cleaning or aligning); *prohibits comm. use* — if the dataset or model's license information includes a non-commercial clause; in *dataset id*, an asterisk (*) indicates that we were unable to identify where the dataset is or was made available; in *dataset name*, some datasets were not given names by their creators.

still available for download. Racial Faces in the Wild [83] also appears available, but requires sending an email to obtain access. Further, we found that the original MS-Celeb-1M dataset, while taken down by Microsoft, continues to be available through third-party sites such as Academic Torrents [24]. We also identified 20 GitHub repositories that continue to make available models pre-trained on MS-Celeb-1M data.

Clearly, one of the goals of retraction is to limit the availability of datasets. Achieving this goal requires addressing all locations where the data might already be or might become available.

Continued use. Besides being available, both MS-Celeb-1M and DukeMTMC have been used in numerous research papers after they were retracted in April 2019. In our sample of papers, we found that DukeMTMC and its derivatives had been used 73 times and MS-Celeb-1M and its derivatives had been used 54 times in 2020. Because our samples are 20% of our entire corpus, this equates to hundreds of uses in total. (See Figure 1 for a comparison of use to previous years.)

This use further highlights the limits of retraction. Many of the cases we identified involved derivatives that were not retracted. Indeed, 72 of 73 DukeMTMC uses were through derivative datasets, 63 of which came from the DukeMTMC-ReID dataset, a derivative that continued to be available. Similarly, only 11 of 54 MS-Celeb-1M uses were through the original dataset, while 17 were through derivative datasets and 26 were through pre-trained models.

One limitation of our analysis is that the use of a dataset in a paper published in 2020 (six months or more after retraction) could mean several things. The research could have been initiated after retraction, with the researchers ignoring the retraction and obtaining the data through a copy or a derivative. The research could have begun before the retraction and the researchers may not have learned of the retraction. Or, the research could already have been under review. Regardless, it is clear that 18 months after the retractions, they have not had the effect that one might have hoped for.

Retractions lacked specificity and clarity. In light of the continued availability and use of both these datasets, it is worth considering whether the retractions included sufficient information about why other researchers should refrain from using the dataset.

After the retraction, the authors of the DukeMTMC dataset issued an apology in *The Chronicle*, Duke's student newspaper, noting that the data collection had violated IRB guidelines in two respects: "Recording outdoors rather than indoors, and making the data available without protections." [80] However, this explanation did not appear on the website that hosted the dataset, which was simply taken down, meaning that not all users looking for the dataset would encounter this information. The retraction of MS-Celeb-1M fared worse: Microsoft never stated ethical motivations for removing the dataset, though the removal followed soon after multiple reports critiquing the dataset for privacy violations [45]. Rather, according to reporting by *The Financial Times*, Microsoft stated that the dataset was taken down "because the research challenge is over" [66]. The website that hosted MS-Celeb-1M is also no longer available. Neither retraction included calls to not use the data.

The disappearance of the websites also means that license information is no longer available through these sites. We were able to locate the DukeMTMC license through GitHub repositories of derivatives. We were unable to locate the MS-Celeb-1M license—which prohibits the redistribution of the dataset or derivatives—except through an archived version.[7] We discuss shortcomings of dataset licenses in Section 5.

We also identified public efforts to access and preserve these datasets, perhaps indicating confusion about the substantive meaning of the dataset's retractions. We found three and two Reddit posts inquiring about the availability of DukeMTMC and MS-Celeb-1M, respectively, following their retraction. Two of these posts (one for each dataset) noted or referred to investigations about potential privacy violations, but still inquired about where the dataset could be found. These posts are listed in Table 5.

In contrast to the retractions of DukeMTMC and MS-Celeb-1M, the retraction of TinyImages was more clear. On the dataset's website, the creators ask that "the community to refrain from using it in future and also delete any existing copies of the dataset that may have been downloaded" [3].

[7] An archived version from April 2017 (found via [45]) is available at `http://web.archive.org/web/20170430224804/http://msceleb.blob.core.windows.net/ms-celeb-v1-split/MSR_LA_Data_MSCeleb_IRC.pdf`.

Table 5: Reddit posts inquiring about how to access DukeMTMC and MS-Celeb-1M after their retractions.

URL	Dataset	Post contents
https://www.reddit.com/r/computervision/comments/drg802/does_anyone_have_dukemtmc_dataset/	DukeMTMC	Hi, currently I am working on my thesis in broad terms "Person re-Identification". All papers that tackle this problem in a way or another this dataset. It was listed as unavailable in the summer of 2019, so I am stuck in the process of finding quality data. Is there any chance that you have or know someone that has this dataset somewhere? Thanks in advance.
https://www.reddit.com/r/datasets/comments/dj6zrh/duke_mtmc_alternative/	DukeMTMC	As of June 2019, the Duke MTMC surveillance dataset was discontinued following a privacy investigation by the financial times. Does anyone know of an alternative source to download it from, or just an alternative dataset all together?
https://www.reddit.com/r/datasets/comments/fpvs47/does_anyone_have_approach_to_get_dukemtmc_dataset/	DukeMTMC	Does anyone have approach to get DukeMTMC dataset? if not, please recommend some other pedestrian datasets in MTMC. Thanks a lot!
https://www.reddit.com/r/datasets/comments/cvq6sa/download_raw_msceleb1m/	MS-Celeb-1M	Hi, I need to download the original MS-Celeb-1M (academic purposes). I tried on megapixels but I could not find any link. There is a reference to a clean dataset (https://github.com/PINTOFSTU/C-MS-Celeb), which, in fact, its only a label list. At academic torrents there is a torrent file, but this dataset is not the original, the images are already croped or aligned. Thanks in advance,
https://www.reddit.com/r/DataHoarder/comments/bxz19f/anyone_have_it_microsoft_takes_down_huge/	MS-Celeb-1M	Anyone have it?: Microsoft takes down huge MS-Celeb-1M facial recognition database

F Supplement: Analysis of license restrictions

We describe findings summarized in Table 3 about the effectiveness of license restrictions for mitigating harms.

Licenses do not effectively restrict production use. We analyzed the licensing information for DukeMTMC, MS-Celeb-1M, and LFW, and determined the implications for production use. Datasets are at a greater risk to do harm in production settings, where characteristics of a dataset directly affect people.

DukeMTMC is released under the CC BY-NC-SA 4.0 license, meaning that users may freely share and adapt the dataset, as long as attribution is given, it is not used for commercial purposes, derivatives are shared under the same license, and no additional restrictions are added to the license. Benjamin et al. [13] note many possible ambiguities in a "non-commercial" designation for a dataset. We emphasize, in particular, that this designation allows the possibility for non-commercial production use. Models deployed by nonprofits and governments maintain risks associated with commercial models. Additionally, there is legal ambiguity regarding whether models trained on the data may be used for commercial purposes.

MS-Celeb-1M is released under a Microsoft Research license agreement,[8] which has several stipulations, including that users may "use and modify this Corpus for the limited purpose of conducting non-commercial research." Again, implications for commercial use of pre-trained models may be ambiguous.

LFW was released without any license. In 2019, a disclaimer was added on the dataset's website, indicating that the dataset "should not be used to conclude that an algorithm is suitable for any commercial purpose" [2]. The lack of an original license meant that the dataset's use was entirely unrestricted until 2019. Furthermore, while it includes useful guiding information, the disclaimer does not hold legal weight. Additionally, through an analysis of results given on the LFW website [2], we found four commercial systems that clearly advertised their performance on the datasets, though we do not know if the disclaimer is intended to discourage this behavior:

- **Innovative Technology.** `https://www.innovative-technology.com/icu`
  "Using our own AI algorithms developed over many years, ICU offers an accurate (99.88%*) precise and affordable facial recognition system *Source: LFW"
- **Oz Forensics.** `https://ozforensics.com/#main_window`
  "The artificial intelligence algorithms recognize people with 99.87% accuracy."
- **IntelliVision.** `https://www.intelli-vision.com/facial-recognition/`
  "Facial recognition accuracy over 99.5% on public standard data sets. It scores the following accuracy in the leading public test databases – LFW: 99.6%, YouTube Faces: 96.5%, MegaFace (with 1000 people/distracters): 95.6%."
- **CyberExtruder.** `https://cyberextruder.com/aureus-insight/`
  "To that end, test results from well-known, publicly-available, industry standard data sets including NIST's FERET and FRGC tests and the UMass Labeled Faces in the Wild data set are shown below."

Because LFW is a relatively small dataset, its use as training data in production settings is unlikely. Risk remains, however, as the use of its performance as a benchmark on commercial systems can lead to overconfidence both among the system creators and potential clients.

ImageNet's "terms of access" specifies that the user may use the database "only for non-commercial research and educational purposes." Again, implications for commercial use of pre-trained models may be ambiguous.

Derivatives do not always inherit original terms. DukeMTMC, MS-Celeb-1M, and ImageNet—according to their licenses—may only be used for non-commercial purposes. We analyzed available derivatives of each dataset to see if they include a non-commercial use designation. All four DukeMTMC derivative datasets included the designation. Four of seven MS-Celeb-1M derivative datasets included the designation. Only three of 21 repositories containing models pre-trained on MS-Celeb-1M included the designation. We also identified 12 repositories containing models pre-trained on ImageNet, of which only three restricted commercial use. Furthermore, Keras, PyTorch, and MXNet all come built in with numerous models pre-trained on ImageNet, and are licensed for commercial use. (This analysis does not apply to LFW, which was released with no license.)

Thus, we found mixed results of license inheritance. We note that DukeMTMC's license specifies that derivatives must include the original license. Meanwhile, MS-Celeb-1M's license, which prohibits derivative distribution in the first place, is no longer publicly accessible, perhaps partially explaining our findings. Licenses are only effective if actively followed and inherited by derivatives.

The loose licenses associated with the pre-trained models are particularly notable. Of the 21 repositories containing models pre-trained on MS-Celeb-1M, seven contained the MIT license, one contained the Apache 2.0 license, and one contained the BSD-2-Clause. Each of these licenses permit commercial use. Additionally, nine repositories were released with no license at all.

[8] The license is no longer publicly available. An archived version is available here: `http://web.archive.org/web/20170430224804/http://msceleb.blob.core.windows.net/ms-celeb-v1-split/MSR_LA_Data_MSCeleb_IRC.pdf`

Table 6: Discussion posts about the legality of the commercial use of models pre-trained on non-commercial data.

Discussion site	Dataset discussed	URL
GitHub	ImageNet	https://github.com/keras-team/keras-applications/issues/140
GitHub	ImageNet	https://github.com/keras-team/keras/issues/13362
GitHub	ImageNet	https://github.com/pytorch/vision/issues/2597
GitHub	ImageNet	https://github.com/tensorflow/models/issues/9131
GitHub	LIP	https://github.com/Engineering-Course/LIP_JPPNet/issues/42
Twitter	ImageNet	https://twitter.com/viglovikov/status/1296292326478761984
Kaggle	ImageNet	https://www.kaggle.com/c/deepfake-detection-challenge/discussion/131121
Reddit	ImageNet	https://www.reddit.com/r/deeplearning/comments/9lalpv/using_pretrained_deep_neural_networks_for/
Reddit	ImageNet	https://www.reddit.com/r/MachineLearning/comments/4eu2vd/can_pretrained_networks_be_used_in_commercial/
Reddit	ImageNet	https://www.reddit.com/r/MachineLearning/comments/7eor11/d_do_the_weights_trained_from_a_dataset_also_come/
Reddit	ImageNet	https://www.reddit.com/r/MachineLearning/comments/id4394/d_is_it_legal_to_use_models_pretrained_on/
Reddit	Scene Flow	https://www.reddit.com/r/MLQuestions/comments/hwxfth/if_i_have_a_dataset_whose_license_restricts_me/
Reddit	FFHQ	https://www.reddit.com/r/MachineLearning/comments/lzmm75/d_legal_mess_in_ml_datasetspretrained_modelscode/
MXNet	ImageNet	https://discuss.mxnet.apache.org/t/commercial-use-license-for-pre-trained-models/3343

G Supplement: Posts discussing legality of pre-trained models

As discussed in Section 5, we identified 14 posts discussing the legality of using models pre-trained on a non-commercial dataset for commercial purposes. We list these posts in Table 6. We identified these posts via four Google searches with the query "pre-trained model commercial use." We then searched the same query on Google with "site:www.reddit.com," "site:www.github.com," "site:www.twitter.com," and "site:www.stackoverflow.com." These are four sites where questions about machine learning are posted. For each search, we examined the top 10 sites presented by Google. Within relevant posts, we also extracted any additional relevant links included in the discussion.

H Supplement: Dataset management and citation

In Section 7, we showed how dataset management and citation can help mitigate harms through facilitating documentation, transparency and accountability, and tracking, and summarized findings showing how current practices fall short in achieving these aims. We present these findings in detail below.

Dataset management practices raise concerns for persistence. Whereas other academic fields utilize shared repositories,[9] machine learning datasets are often managed through the websites of individual researchers or academic groups. None of the 38 datasets in our analysis are managed through shared repositories. Unsurprisingly, we found that some datasets were no longer maintained (which is different from being retracted).

We were only able to find information about D31 and D38 through archived versions of sites found via the Wayback Machine. And even after examining archived sites, we were unable to locate information about D6, D17, and D25. Another consequence is the lack of persistence of documentation. Ideally, information about a dataset should remain available even if the dataset itself is no longer available. But we found that after DukeMTMC and MS-Celeb-1M were taken down, so too were the sites that contained their terms of use.

Dataset references can be difficult to disambiguate. Clear dataset citation is important for harm mitigation. However, datasets are not typically designated as independent citable research objects like academic papers are. This is evidenced by a lack of standardized permanent identifiers, such as

[9] *Nature* provides specific guidance for both field-specific and general data repositories (`https://www.nature.com/sdata/policies/repositories`).

Table 7: Examples of dataset references that were challenging to disambiguate.

Reference	Attempted disambiguation
"Experiments were performed on four of the largest ReID benchmarks, i.e., Market1501 [45], CUHK03 [17], DukeMTMC [33], and MSMT17 [40] … DukeMTMC provides 16,522 bounding boxes of 702 identities for training and 17,661 for testing."	Here, the dataset is called DukeMTMC and the citation [33] is of DukeMTMC's associated paper. However, the dataset is described as an ReID benchmark. Moreover, the statistics given exactly match the popular DukeMTMC-ReID derivative (an ReID benchmark). This leads us to believe DukeMTMC-ReID was used.
"We used the publicly available database Labeled Faces in the Wild (LFW)[6] for the task. The LFW database provides aligned face images with ground truth including age, gender, and ethnicity labels."	The name and reference both point to the original LFW dataset. However, the dataset is described to contain aligned images with age, gender, and ethnicity labels. The original dataset contains neither aligned images nor any of these annotations. There are, however, many derivatives with aligned versions or annotations by age, gender, and ethnicity. Since no other description was given, we were unable to disambiguate.
"MS-Celeb-1M includes 1M images of 100K subjects. Since it contains many labeling noise, we use a cleaned version of MS-Celeb-1M [16]."	The paper uses a "cleaned version of MS-Celeb-1M," but the particular one is not specified. (There are many cleaned versions of the dataset.) The citation [16] is to the original MS-Celeb-1M's associated paper and no further description is given. Therefore, we were unable to disambiguate.

DOIs. None of the 38 datasets we encountered in our analysis had such identifiers. Datasets are often assigned DOIs when added to shared data repositories.

Without dataset-specific identifiers, we found that datasets were typically cited with a combination of the dataset's name, a description, and paper citations. In many cases, an associated paper is cited—a paper through which a dataset was introduced or that the dataset's creators request be cited. In some cases, a dataset does not have a clear associated paper. For example, D31 was not introduced in an academic paper and D20's creators suggest three distinct academic papers that may be cited. This practice can lead to challenges in identifying and accessing the dataset(s) used in a paper, especially when the name, description, or citation conflict. There is a discrepancy between the roles of citation for attribution and documentation: providing sufficient attribution does not necessarily imply that sufficient documentation is given, and vice versa.

In our analysis, 42 papers included dataset references that we were unable to fully disambiguate. Oftentimes, this was a result of conflating a dataset with its derivatives. For example, we found nine papers that suggested that images in LFW were annotated with attributes or keypoints, but did not specify where these annotations were obtained. (LFW only contains images labeled with identities and many derivatives of LFW include annotations.) Similarly, seven papers indicated that they used a cleaned version of MS-Celeb-1M, but did not identify the particular derivative. We were able to disambiguate the references in 404 papers using a dataset or a derivative, but in many of these instances, making a determination was not direct (for instance, see the first example in Table 7).

Datasets and documentation are not directly accessible from citations. We found that accessing datasets from papers is not currently straightforward. While data access requirements, such as sections dedicated to specifying where datasets and other supporting materials may be found, are common in other fields, they are rare in machine learning. We sampled 60 papers from our sample that used DukeMTMC, MS-Celeb-1M, LFW, or one of their derivative datasets, and only six provided access information (each as a URL).

Furthermore, the descriptors we mentioned above—paper citations, name, and description—do not offer a direct path to the dataset. The name of a dataset can sometimes be used to locate the dataset via web search, but this works poorly in many instances—for example, when a dataset is not always associated with a particular name or when the dataset is not even available. Datasets D27, D28, D31, D32, and D38 are not named. In other cases, datasets may be known by multiple names. Equating datasets can be challenging. As one GitHub user commented: "Since there are many different names regarding different versions of ms1m dataset, below is my own understanding for these different names: ms1m-v1 = ms1m-ibug[,] ms1m-v2 = ms1m-arcface[,] both of them are detected by mtcnn

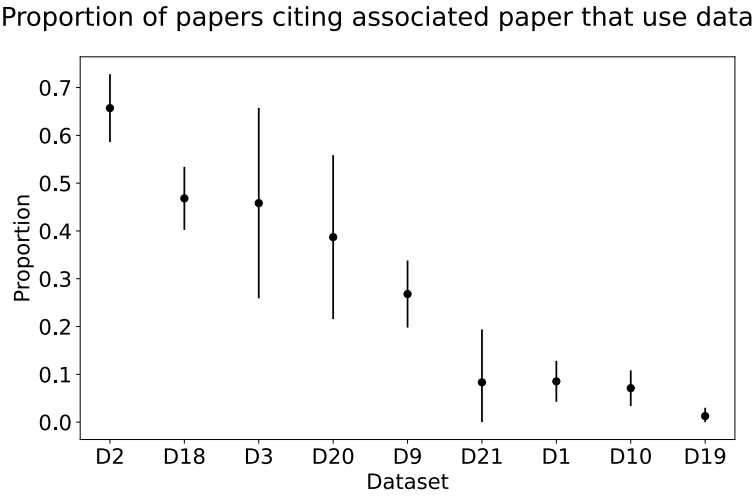

Figure 3: Papers citing associated papers often do not use the associated dataset. The proportion that do varies greatly across different datasets. Here, we include associated papers for which we sampled at least 20 citing papers, and show 95 percent confidence intervals.

and use the same alignment procedure. Am I understanding correctly?"[10]. Here, note that "ms1m" is a common abbreviation for MS-Celeb-1M.

Citations of an associated paper also do not directly convey access information. As an alternate strategy, we were able to locate some datasets by searching for personal websites of the dataset creators or of their associated academic groups. However, as we mentioned earlier, we were still unable to locate D6, D17, and D25, even after looking at archived versions of sites.

Current infrastructure makes tracking dataset use difficult. A lack of dataset citations also makes it difficult to track dataset use. Citations of associated papers are not necessarily effective proxies in this respect. On one hand, the proportion of papers citing an associated paper that use the corresponding dataset varies significantly (see Figure 3). This is because papers citing an associated paper may be referencing other ideas mentioned by the paper. On the other hand, some datasets may be commonly used in papers that do not cite a particular associated paper. Of the papers we found to use DukeMTMC-ReID, 29 cited the original dataset, 63 cited the derivative dataset, and 50 cited both. Furthermore, some datasets may not have a clear associated paper, and various implementations of pre-trained models are unlikely to have associated papers. Thus, associated papers—as currently used—are an exceedingly clumsy way to track the use of datasets.

Tracking derivative creation presents an even greater challenge. Currently, there is no clear way to identify all derivatives of a dataset. The websites of LFW and DukeMTMC (the latter no longer online), maintained lists of derivatives. However, our analysis reveals that these lists are far from complete. Proponents of dataset citation have suggested the inclusion of metadata indicating provenance in a structured way (thus linking a dataset to its derivatives) [39], but such a measure has not been adopted by the machine learning community.

Ambiguities in dataset citation and the instability of datasets present fundamental challenges to alternative approaches to automating the tracking of dataset use and derivative creation. Meanwhile, the adoption of standard practices in dataset management and citation can enable both of these tasks.

[10] `https://github.com/deepinsight/insightface/issues/513` Another post expressing similar confusion about MS-Celeb-1M derivatives is available here: `https://github.com/deepinsight/insightface/issues/566`.

Table 8: GitHub repositories of models pre-trained on MS-Celeb-1M

Model class	Still available	License	GitHub url
M1	✓	MIT	https://github.com/cydonia999/VGGFace2-pytorch
M1		none	https://github.com/ox-vgg/vgg_face2
M2	✓	MIT	https://github.com/seasonSH/Probabilistic-Face-Embeddings
M3	✓	non-commercial research	https://github.com/deepinsight/insightface/tree/master/recognition/arcface_torch
M3	✓	MIT	https://github.com/auroua/InsightFace_TF
M3	✓	MIT	https://github.com/AIInAi/tf-insightface
M3	✓	MIT	https://github.com/TreB1eN/InsightFace_Pytorchh
M3	✓	none	https://github.com/ronghuaiyang/arcface-pytorch
M3	✓	none	https://github.com/gehaocool/CombinedMargin-caffe
M3	✓	none	https://github.com/luckycallor/InsightFace-tensorflow
M3	✓	MIT	https://github.com/wang-xinyu/tensorrtx
M3	✓	none	https://github.com/deepinsight/insightface/wiki/Model-Zoo
M3	✓	Apache 2.0	https://github.com/foamliu/InsightFace-PyTorch
M3	✓	non-commercial research	https://github.com/leondgarse/Keras_insightface
M4	✓	none	https://github.com/AlfredXiangWu/LightCNN
M4	✓	non-commercial	https://github.com/AlfredXiangWu/face_verification_experiment
M4	✓	none	https://github.com/yxu0611/Tensorflow-implementation-of-LCNN
M4	✓	none	https://github.com/lyatdawn/LightCNN-mxnet
M5	✓	MIT	https://github.com/davidsandberg/facenet/tree/accd6881d58b3bf7bfbdc12bae2d6dde738ba48e
M5	✓	none	https://github.com/nyoki-mtl/keras-facenet
M6	✓	BSD-2-Clause	https://github.com/penincillin/DREAM

Table 9: GitHub repositories of models pre-trained on ImageNet

License	GitHub url
none	https://github.com/PengBoXiangShang/MobileNetV3_PyTorch
none	https://github.com/qubvel/resnet_152
none	https://github.com/visionNoob/pytorch-darknet19
MIT	https://github.com/flyyufelix/DenseNet-Keras
MIT	https://github.com/qubvel/segmentation_models.pytorch
MIT	https://github.com/Alibaba-MIIL/ImageNet21K
non-commercial research	https://github.com/HiKapok/TF-SENet
non-commercial research	https://github.com/HiKapok/Xception_Tensorflow
BSD-3-Clause	https://github.com/Cadene/pretrained-models.pytorch
Apache-2.0	https://github.com/kuan-wang/pytorch-mobilenet-v3
Apache-2.0	https://github.com/pudae/tensorflow-densenet
Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International Public License for Noncommercial use only	https://github.com/Res2Net/Res2Net-PretrainedModels

I Supplement: Identifying models pre-trained on MS-Celeb-1M and ImageNet

In our analysis, we identified GitHub repositories containing models pre-trained on MS-Celeb-1M and ImageNet. We describe our methodology below.

For each of the six pre-trained "model classes" we identified in our analysis (all pre-trained on MS-Celeb-1M), we identified if and where the corresponding pre-trained models are available on GitHub. We first identified repositories linked in the papers using the model. Within these repositories, we also examined any linked third-party implementations. We further searched the name of the model class on GitHub and examined the first 10 results for if they contained a model pre-trained on MS-Celeb-1M. This resulted in a total of 21 repositories (listed in Table 8), 20 of which currently contain a model pre-trained on MS-Celeb-1M.

We performed a similar search to identify models pre-trained on ImageNet. For this, we searched "ImageNet pretrained" on GitHub and then examined the first 20 results. This yielded 12 repositories that contained a model pre-trained on ImageNet, listed in Table 9.

J Recommendations: The role of the IRB

In Section 8, we outlined recommendations for several stakeholders. In particular, we suggested that PCs take a larger role in regulating dataset use and creation. Here, we address the role of Institutional Review Boards (IRBs), which have historically played a fundamental role in regulating research ethics.

Researchers have recently called for greater IRB oversight in dataset creation [70], and IRBs have certain natural advantages in regulating datasets. IRBs may have more ethics expertise than program committees; IRBs are also able to review datasets prior to their creation. Thus, IRBs can prevent harms that occur during the creation process.

However, conceived first to address biomedical research, IRBs have been an imperfect fit for data-centered research. Notably "human subjects research" has a narrow definition and thus many of the datasets (and associated research) that have caused ethical concern in machine learning would not fall under the purview of IRBs. An even more significant limitation is that IRBs are not allowed to consider downstream harms [61].[11]

Unless and until the situation changes, our primary recommendation regarding IRBs is for researchers to recognize that research being approved by the IRB does not mean that it is "ethical," and for IRBs themselves to make this as clear as possible.

[11] "The IRB should not consider possible long-range effects of applying knowledge gained in the research (e.g., the possible effects of the research on public policy) as among those research risks that fall within the purview of its responsibility" (45 CFR §46.111).